#6

KONINKRIJK DER NEDERLANDEN

Bureau voor de Industriële Eigendom

Hierbij wordt verklaard, dat in Nederland op 25 mei 1999 onder nummer 1012148,

ten name van:

KONINKLIJKE KPN N.V.

te Groningen

een aanvrage om octrooi werd ingediend voor:

"Spraakverwerkend systeem",

en dat de hieraan gehechte stukken overeenstemmen met de oorspronkelijk ingediende stukken.

Rijswijk, 25 januari 2000.

De Directeur van het Bureau voor de Industriële Eigendom,
voor deze,

P.J.C. van den Nieuwenhuijsen.

Application: 10/019,766
Confirmation No.: 1446
Docket No.: PTT-127(402568US)
Call: Peter L. Michaelson, Esq.
(732)530-6671

UITTREKSEL

Om de performance van spraakherkenning onder mobiele omstandigheden te verbeteren, is het gebruikelijk dat men spraakmateriaal verzamelt teneinde nauwkeuriger modellen van de spraak te kunnen maken. Echter, met enige regelmaat wordt de errorcorrectie veranderd door de fabrikant, waardoor de mismatch tussen training en realiteit toeneemt. Bovendien worden transmissiefouten momenteel 'opgevangen' door ze mee te nemen in het trainingsproces, hetgeen de kans vergroot op 'garbage-in, garbage-out'. Teneinde deze nadelen te ondervangen wordt de informatie die downstream (1, 2) in de frames beschikbaar is over de framekwaliteit (BFI) en de aanwezigheid van spraak (SP), gebruikt om de upstream spraakherkenner (20) dynamisch te besturen. Het resultaat is dat van niet-correct veronderstelde frames alleen het correcte deel gebruikt wordt, en frames waarin geen spraak verstuurd is, maar waarin sprake is van stilte, door de spraakherkennr worden genegeerd.

(FIG. 1)

Spraakverwerkend systeem

ACHTERGROND VAN DE UITVINDING

De uitvinding heeft betrekking op een spraakverwerkend systeem, omvattende spraakherkenningsmiddelen voor de verwerking van vanuit een bron aan een spraakingang van dat spraakverwerkende systeem toegevoerd signaal (DATA). Bekend is dat de kwaliteit van spraakherkenning aan de ontvangstzijde van bijv. een GSM verbinding momenteel onvoldoende is. Als de herkenner zich in het netwerk bevindt, wordt het herkenresultaat op het ontvangen en gedecodeerde GSM spraaksignaal mede beïnvloed door de hoeveelheid artificieel gegenereerde ruis die op basis van aan zendzijde gedetecteerde stilte wordt toegevoegd en de ontvangen ruis en verstoringen die het gevolg zijn van gedecodeerde transmissie fouten op het radiopad. Om de herkenning te verbeteren, is het gebruikelijk spraakmateriaal te verzamelen dat via GSM verzonden is geweest en dat materiaal te gebruiken om nieuwe spraakmodellen te ontwikkelen, die getraind zijn op spraaksignalen die (artificieel gegenereerde) ruis en distorties door transmissiefouten bevatten, waardoor de mismatch tussen trainsituatie en de herkenrealiteit verkleind kan worden.

Het bekende heeft de volgende nadelen: de performance van de spraakherkenner is door het trainen op de ontvangen en gedecodeerde spraaksignalen slechts beperkt te verbeteren omdat:

1) het decoderen van bijv. gecodeerde GSM signalen niet gestandaardiseerd is (alleen het encoderen is gestandaardiseerd), wat betekent dat er in de praktijk situaties ontstaan waarin de spraakherkenner getraind is op een andere GSM spraakdecoder dan aan de input van de herkenner wordt toegepast. Bijvoorbeeld de error- correctie die wordt toegepast in de decoder wordt regelmatig veranderd omdat de fabrikant een betere manier heeft

gevonden om transmissiefouten (waardoor beschadigde spraak ontstaat) zodanig te bewerken dat een groot deel van deze fouten verborgen wordt (en dus niet of nauwelijks merkbaar voor het menselijk gehoor). Dit heeft tot gevolg dat er een mismatch ontstaat tussen de trainingset waarop de spraakmodellen zijn gebaseerd en de werkelijke spraak.
2) men door te trainen op spraak met transmissiefouten weliswaar de fouten reeds modelleert in de spraakmodellen (die daardoor complexer worden), maar het is niet gegarandeerd dat de algehele kwaliteit van de herkenning toeneemt, want vaak geldt: garbage-in, garbage-out.
3) niet vooraf bekend is of een signaal spraak of stilte (vanaf de zendzijde) bevat. Omdat aan de ontvangstzijde artificieel gegeneerde ruis wordt toegevoegd (comfort noise) wanneer er stiltes geconstateerd zijn, daalt de performance van de spraakherkenning omdat de herkenner zal proberen de ruis te 'herkennen'.

SAMENVATTING VAN DE UITVINDING

De uitvinding beoogt de genoemde nadelen te ondervangen en de performance te verbeteren van automatische spraakherkensystemen die opereren aan de ontvangstzijde van een spraakframe georiënteerde telefonische spraakverbinding. Dit kan zijn bijv. GSM, UMTS of Voice Over IP. De kern van de uitvinding is dat aan ontvangstzijde niet alleen een spraaksignaal aan het spraakherkensysteem wordt aangeboden, maar ook signaalparameters die informatie geven over karakteristieken van het ontvangen signaal.
Bijvoorbeeld betreft het parameters die duiden op de aan- of afwezigheid van spraakenergie in het ontvangen signaal of op de betrouwbaarheid van het ontvangen signaal blijkens aan zendzijde toegevoegde redundancy checks (bijv. CRC's). Bij GSM worden dergelijke parameters op basis van frames berekend. De in het kader van de uitvinding van belang

zijnde parameters zijn daar ondermeer de BFI (Bad Frame Indicator), bijv. berekend uit de CRC waarden per frame, en de SID (SIlence Descriptor) afgeleid van een parameter SP (Speech Flag). Deze parameters worden in GSM tot dusverre alleen gebruikt voor detectie van fouten in de ontvangen spraakframes resp. voor zenderbesturing (alleen zenden bij de aanwezigheid van spraak).

Besturing van een spraakherkenner door klassificerende parameters bevordert de accuraatheid van de herkenning doordat artificieel gegeneerde ruis genegeerd kan worden, en kapotte frames hetzij genegeerd worden, hetzij aangepast, bijvoorbeeld partieel, verwerkt worden. Behalve de bovengenoemde parameters, de BFI en SID, wordt ook gebruik gemaakt van een "coding mode" parameter die de betekenis van de spraakframe bits definieert (FR, EFR, of de verschillende modes waarin AMR kan werken). Aan de hand hiervan wordt het in de spraakherkenner werkzame herkenalgoritme aangepast aan de karakteristieken waarmee het spraaksignaal is gecodeerd en gedecodeerd.

FIGUURBESCHRIJVING

De werking van de uitvinding wordt aan de hand van enige figuren nader toegelicht. Als voorbeeld nemen we het huidige deel van het GSM systeem dat gebruik maakt van een Enhanced Full Rate (EFR) codec. Hetzelfde geldt echter voor een Full Rate (FR) codec, en voor de (toekomstige) Adaptive Multi Rate codec (AMR). Figuur 1 toont twee terminals -een eerste, mobiele terminal zoals een GSM handset, en een tweede, vaste terminal zoals een GSM basisstation- die met elkaar kunnen communiceren via een draadloos medium 9. In de figuur wordt alleen upstream communicatie -van handset naar basisstation- voorgesteld.

De in het bovenste deel van figuur 1 getoonde handset omvat twee modules of subsystemen, te weten een TX/DTX Handler 1 (DTX staat voor Discontinous Transmission) en een TX Radio

Subsystem 2. Module 1 omvat een microfoon 3, een spraak-encoder 4 en een Voice Activity Detector (VAD) 5. Module 2 omvat een kanaal-encoder 6, een Speech Flag monitor 7 en een zender 8. Door de microfoon 3 ontvangen signalen worden toegevoerd aan zowel de spraak-encoder 4 als naar de VAD 5. In de VAD 5 wordt gedetecteerd of de microfoon 3 spraak of stilte opvangt. Dit wordt gecodeerd met een "Speech Flag" (SP), welke wordt meegestuurd in elk spraakframe. In de kanaal-encoder 6 wordt het in encoder 4 gecodeerde microfoon-signaal gecodeerd tot via zender 8 verzendbare frames. Aan de frames is wordt redundante informatie toegevoegd, zoals een checksum code (CRC) aan de hand waarvan aan ontvangzijde kan worden berekend of het frame correct is overgedragen. In bepaalde gevallen kan een niet-correct overgedragen frame met behulp van deze redundante informatie worden gecorrigeerd.

Tijdens de opbouw van de verbinding wordt vastgesteld welk codeeralgoritme gebruikt wordt, hetgeen gerepresenteerd kan worden als de parameter CM ("coding mode"). Bij bepaalde spraakcodecs (bijv. AMR) wordt de "coding mode"-parameter per frame meegestuurd en wordt de herkenner hiermee dynamisch aangestuurd. Bij andere spraakcodecs wordt de parameter eenmalig, aan het begin van een sessie, naar de ontvangzijde overgedragen.

Aldus zendt zender 8 een frame-gecodeerd signaal uit dat data (het eigenlijke signaal), de parameter SP, de parameter CM (bij bepaalde spraakcodecs) en redundante informatie, zoals de checksum CRC bevat.

De ontvangende terminal, onderaan in figuur 1, omvat twee modules of subsystemen in een GSM basisstation, te weten een RX Radio System 11, de tegenhanger van module 2 van de handset, en een RX DTX Handler 12, de tegenhanger van module 1. Module 11 omvat een ontvanger 13, een kanaal-decoderings- en foutcorrectiemodule 14 en een

parameterdetector 15; die laatste detecteert de aanwezigheid en de waarde van de met het datasignaal meegezonden parameter SP en, indien aanwezig, de parameter CM. Module 12 omvat een spraak-decoder 16 en een verdere verwerkingsmodule 17.

De ingang van een spraakherkenmodule 20 is -overigens op zich conform de stand van de techniek- aangesloten op de uitgang van de kanaal-decoder 14. De spraakherkenner 20 bewerkt dus het nog niet spraak-gedecodeerde datasignaal (spraak). Conform de onderhavige uitvinding wordt de spraakherkenner 20 aangestuurd door één of meer signaalparameters die via detector 15 worden ontvangen. De basis van de parameter SP wordt aan zendzijde, in de GSM handset, gevormd, los van de signaal-inhoud van het ontvangen datasignaal. In de foutcorrectiemodule 14 worden de ontvangen frames voorafgaand aan decodering op correctheid onderzocht aan de hand van de meegezonden redundante informatie. Niet-correcte frames worden als zodanig aangemerkt of zo mogelijk hersteld (in simpele gevallen). Correcte frames worden doorgegeven naar de spraakdecoder 15. Wanneer een frame niet gecorrigeerd kan worden, geeft module 14 een BFI ("Bad Frame Indicator") parameter af aan detectormodule 15. Volgens de uitvinding wordt die BFI, behalve aan de spraak-decoder 16, eveneens doorgegeven aan de spraakherkenner 20. Op ontvangst van die BFI negeert de spraakherkenner 20 de aangeboden input, of probeert het deel van het frame dat nog wel als correct kan worden aangemerkt (hoewel de BFI gezet is) alsnog te herkennen. De waarde van de BFI parameter werkt met ander woorden als besturingsparameter voor de spraakherkenner, waardoor die alleen correcte frames in één keer bewerkt. Van als kapot aangemerkte frames wordt geprobeerd alleen dat deel te gebruiken dat nog correct is, en als geheel incorrect aangemerkte frames worden genegeerd. Dat bij een

gezette BFI vlag nog steeds een deel van het frame correct kan zijn, komt doordat de bits in de spraakframes in verschillende klassen zijn opgedeeld (in GSM: 1A, 1B en 2). Niet elke klasse wordt op dezelfde manier 'beschermd' door toegevoegde redundante informatie. Bij bijv. GSM geldt dat indien klasse 1A bits als 'beschadigd' worden gekenmerkt (op basis van de CRC), de BFI vlag gezet wordt (sommige fabrikanten zetten deze vlag ook bij beschadigde 1B bits). Dit hoeft echter niet te betekenen dat alle overige bits ook beschadigd zijn. De herkenner neemt als input feature vectoren (Rabiner & Juang, 1993). Elk spraakframe wordt omgezet in een feature vector. De waarden van het deel van het spraakframe dat niet beschadigd is, kunnen nog steeds aangeboden worden aan de herkenner. Dit kan bijvoorbeeld gerealiseerd worden door de gecorrumpeerde features in de feature vectoren één specifieke waarde te geven welke resulteert in een nihil effect op de score van het ontvangen signaal (de Veth, Cranen & Boves, 1998), of door het complete frame te negeren (Lippman & Carlson, 1997). Op ongeveer dezelfde wijze werkt de SID parameter op de werking van de spraakherkenner 20. De SID parameter wordt afgeleid van de waarde van de Speech Flag, zoals die wordt afgegeven door de Voice Activity Detector 5 en verzonden door zender 8. Bij spraak krijgt de SP een bepaalde waarde, en evenzo de SID; bij ontbreken van spraak (stilte) krijgen de SP en daardoor de SID parameter een andere waarde. Het resultaat is dat de spraakherkenner "enabled" is bij de overdracht van een werkelijk spraaksignaal en "disabled" bij de afwezigheid van spraak. Tenslotte is het, zoals hierboven werd aangegeven, mogelijk om de werking van spraakherkenner 20 in te stellen in afhankelijkheid van het codeeralgoritme van de spraak-encoder 4 (bijv. FR, EFR, AMR, etc.). In de figuur geschiedt dat door de middels

hand-shake (dus tijdens de verbindingsopbouw) vastgestelde, of door de per spraakframe meegestuurde parameter CM.

REFERENTIES

Lippmann, R. P. , Carlson, B. A. , "Missing feature theory to actively select features for robust speech recognition with interruptions, filtering and noise", Proc. Of Eurospeech97, Rhodos, Griekenland, 1997.

Rabiner, L. , Juang, B. H. , "Fundamentals of Speech Recognition", Prentice-Hall, Inc. New Jersey, 1993.

Veth, J. de, Cranen, B. , Boves, L. (1998), "Acoustic backing-off in the local distance computation for robust automatic speech recognition", Proc. Of ICSLP 1998, Sydney, Australie.

CONCLUSIES

1. Spraakverwerkend systeem, omvattende spraakherkenningsmiddelen (20) voor de verwerking van een vanuit een bron (1, 2) aan een spraakingang toegevoerd signaal (DATA), GEKENMERKT DOOR middelen voor het beïnvloeden van de werking van de spraakherkenningsmiddelen door één of meer via een besturingsingang toegevoerde besturingsparameters (CM, SID, BFI), waarbij elke besturingsparameter betrekking heeft op een bepaalde karakteristiek van het vanuit de bron aan de spraakherkenningsmiddelen toegevoerde signaal (DATA).

2. Spraakverwerkend systeem volgens conclusie 1, MET HET KENMERK DAT een eerste besturingsparameter (BFI) betrekking heeft op de betrouwbaarheid of correctheid van het toegevoerde signaal en de werking van de spraakherkenningsmiddelen (20) aangepast wordt aan de door die eerste besturingsparameter aangegeven betrouwbaarheid respectievelijk correctheid van het toegevoerde signaal.

3. Spraakverwerkend systeem volgens conclusie 1, MET HET KENMERK DAT een tweede besturingsparameter (SID) betrekking heeft op de spraak/ruis-ratio en de werking van de spraakherkenningsmiddelen (20) aangepast wordt aan de door die tweede besturingsparameter aangegeven spraak/ruis-ratio van het toegevoerde signaal.

4. Spraakverwerkend systeem volgens conclusie 1, waarbij het aan de spraakherkenningsmiddelen (20) toegevoerde signaal in spraakcodeermiddelen (4) aan de bron gecodeerd is, MET HET KENMERK DAT een derde besturingsparameter (CM) betrekking heeft op de modus van spraakcodering in de spraakcodeermiddelen, waarbij de werking van de spraakherkenningsmiddelen (20) aangepast wordt aan de door die derde besturingsparameter aangegeven spraakcoderingmodus.

5. Telecommunicatiesysteem, omvattende een eerste terminal (1, 2) met spraak- en kanaal-encodeermiddelen (4, 6), een

transmissiemedium (9) en een tweede terminal (11, 12) met kanaal- en spraakdecodeermiddelen (13, 16) en een spraakverwerkend systeem volgens conclusie 1, waarbij het genoemde signaal (DATA) vanuit de eerste terminal, via het transmissiemedium aan de spraakingang van de spraakherkenner van de tweede terminal wordt aangeboden, en waarbij elke besturingsparameter (CM, SID, BFI) vanuit de eerste terminal, via het transmissiemedium aan de daartoe bestemde besturingsingang van het spraakverwerkende systeem van de tweede terminal wordt aangeboden.

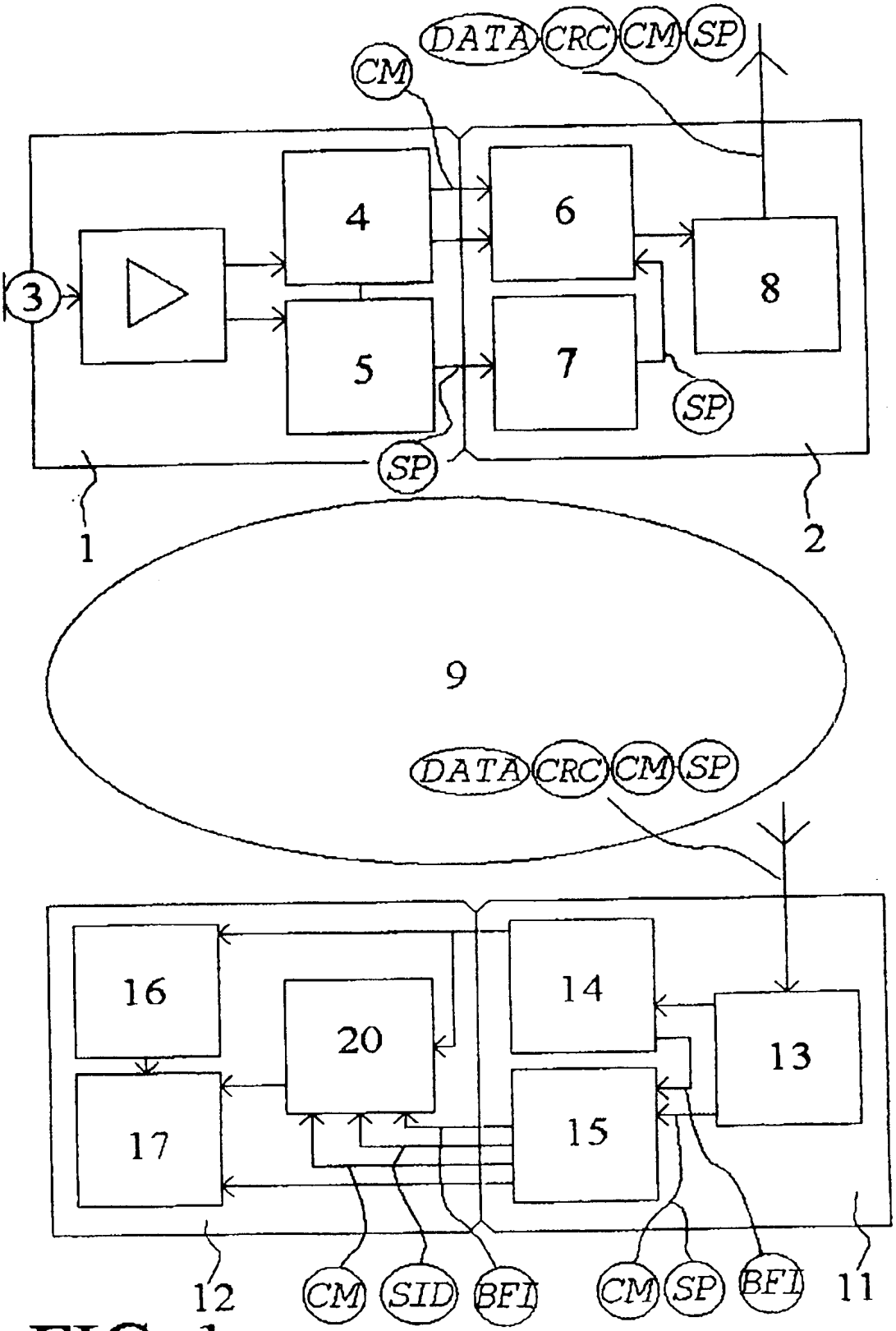
DATA
CRC
CM
SP
CM
4
6
8
3
5
7
SP
SP
1
2
9
DATA
CRC
CM
SP
16
20
14
13
17
15
12
CM
SID
BFI
CM
SP
BFI
11

FIG. 1

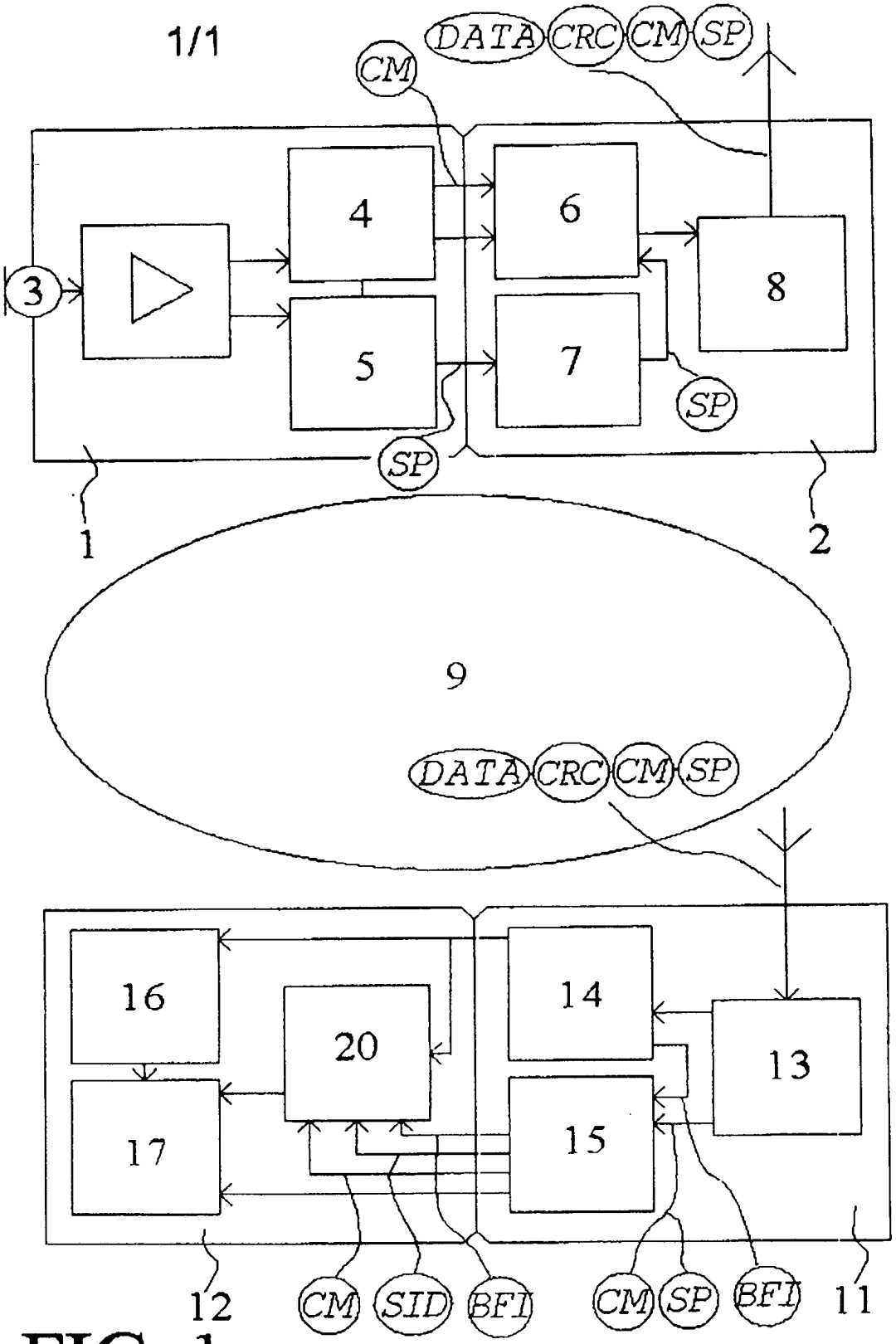
DATA
CRC
CM
SP
CM
4
6
8
3
5
7
SP
SP
1
2
9
DATA
CRC
CM
SP
16
20
14
13
17
15
12
CM
SID
BFI
CM
SP
BFI
11

FIG. 1